Airline ticket category classification

|  |  |  |
| --- | --- | --- |
| Section | Id | Name |
| Sec 2 | 20191700169 | إياد عبده محمد اسماعيل |
| Sec 2 | 20191700168 | إيهاب محمد عبدالمنعم محمد |
| Sec 5 | 20191700127 | الفت ناصر على محمد |
| Sec 2 | 20191700146 | انس احمد محمد عبدالعزيز |
| Sec 2 | 20191700197 | جمال احمد عبدالحكم |

Ada Boost Hyperparameters Combinations  
  
base\_estimator: The base estimator from which the boosted ensemble is built.

n\_estimators: The maximum number of estimators at which boosting is terminated.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| base\_estimator | n\_estimators | Train Time | Train Accuracy | Test Accuracy |
| ExtraTreesClassifier() | 50 | 20 | 0.999 | 0.936 |
| 100 | 37 | 0.999 | 0.936 |
| 200 | 74 | 0.999 | 0.941 |
| DecisionTreeClassifier() | 50 | 55 | 0.999 | 0.9578 |
| 100 | 112 | 0.999 | 0.9572 |
| 200 | 228 | 0.999 | 0.9568 |
| LogisticRegression() | 50 | 95 | 0.4278 | 0.4276 |
| 100 | 196 | 0.4278 | 0.4276 |
| 200 | - | - | - |

Decision Tree Hyperparameters Combinations  
  
Criterion: entropy: splits a node such that it gives the most information gain

min\_samples\_split: minimum number of samples required to split an internal node  
  
min\_samples\_leaf: minimum number of samples required to create a new leaf

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| min\_samples\_split | min\_samples\_leaf | Train Time | Train Accuracy | Test Accuracy |
| 25 | 10 | 1.5 | 0.958 | 0.945 |
| 15 | 1.6 | 0.953 | 0.942 |
| 20 | 1.5 | 0.947 | 0.935 |
| 30 | 10 | 1.4 | 0.947 | 0.935 |
| 15 | 1.4 | 0.957 | 0.944 |
| 20 | 1.5 | 0.953 | 0.942 |
| 40 | 10 | 1.4 | 0.947 | 0.935 |
| 15 | 1.4 | 0.95 | 0.94 |
| 20 | 1.5 | 0.947 | 0.936 |

Logistic Regression Hyperparameters Combinations  
  
solver: algorithm used in the optimization problems  
  
max\_iter: maximum number of iterations taken for the solvers to converge

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| solver | max\_iter | Train Time | Train Accuracy | Test Accuracy |
| liblinear | 50 | 1.57 | 0.4278 | 0.4277 |
| 100 | 1.63 | 0.4278 | 0.4277 |
| 200 | 1.6 | 0.4278 | 0.4277 |
| sag | 50 | 6 | 0.4278 | 0.4277 |
| 100 | 13.5 | 0.4278 | 0.4277 |
| 200 | 26.4 | 0.4278 | 0.4277 |
| saga | 50 | 7 | 0.4278 | 0.4277 |
| 100 | 13 | 0.4278 | 0.4277 |
| 200 | 24.5 | 0.4278 | 0.4277 |

Conclusions  
  
- As we started this phase, we thought that the Decision Tree Classifier would give the best test accuracy  
  
- Although it didn’t show a big difference, both Decision Tree and Ada Boost achieved high accuracy  
  
- However, The Ada Boost Classifier took too long to learn the data while Decision Tree took almost a second  
  
- Both seemed not to overfit and to achieve higher accuracy  
  
- Logistic Regression seemed to underfit as it achieved low train and test accuracy  
  
- The reason for underfitting (Failure) in Logistic Regression is that the features cannot be linearly separated